

# 非母語での発話における使用語彙と通じやすさの関係

和泉 絵美 内元 清貴 井佐原 均

独立行政法人情報通信研究機構 自然言語グループ

E-mail: {emi,uchimoto,isahara}@nict.go.jp

## 1. はじめに

### 1.1. 研究の背景

言語コミュニケーションにおいて、発話意図に沿った適切な語彙を使用することは重要である。特に、非母語でのコミュニケーションにおいては、完璧な文法の運用ができない分、より語彙（特に内容語）に頼らざるを得ないことも多い。また、単語単位での使用だけではなく、複数の単語を適切な組み合わせで使用すること（コロケーションの使用）が、発話をより自然にするための鍵となる。

言語コミュニケーションの第一目的は、メッセージの伝達と理解であるため、聞き手に通じやすい発話を行う必要がある。文や談話の構造、語彙の選び方によって発話の通じやすさは変わる。更に、母語での発話と比べて誤りが含まれる可能性が高い非母語での発話においては、誤りの種類や頻度、出現位置によって発話の通じやすさはより激しく左右される。発話の意味内容を大きく担う語彙に関する誤りが、通じやすさにどのような影響を及ぼすのか、大変興味深いところである。

### 1.2. 研究の目的

言語コミュニケーションにおける語彙の役割は極めて重要であり、同時にコミュニケーション能力(communicative competence)の育成に重きを置く外国語学習(Canale and Swain, 1980)においても、語彙の習得は最重要課題のひとつと言える。

本研究では、非母語での発話における単語レベルの語彙誤りが、発話の通じやすさに及ぼす影響について調査する。まず、日本語母語話者の英語発話データに付与された誤り情報（エラータグ）と英語母語話者によって判定された各文の通じやすさレベルの情報を利用して、誤りの種類が発話の通じやすさに及ぼす影響に関して分析を行う。その結果を元に、語彙の誤りが発話の通じやすさに及ぼす影響について更に調査する。具体的には、まず非母語での言語産出における語彙の誤りパタンの細分類について検討する。その体系に基づいてデータ内の語彙誤りを分類し、各誤りパターンが異なる通じやすさレベルのデータ内でどのように分布しているのかを観察する。

本章以降では、まず2.で日本人英語の通じやすさに関する調査について述べる。次に、3.では、2.の調査結果を踏まえた上で、語彙の誤りパタン

と通じやすさの関係に関する調査について報告する。この中で、非母語での言語産出における語彙誤りパタンの細分類についても詳しく述べる。最後に4.でまとめと今後の課題について述べる。

## 2. 日本人英語の通じやすさに関する調査

どのような発話に通じやすい（にくい）のか、すでに経験的・直感的な感想を持っている学習者もいるが、より明示的に日本人英語の通じやすさを記述することは、非母語でコミュニケーションを行う際の手助けとなると考え、日本人英語の実データを元に調査を行った。

### 2.1. 英語母語話者による通じやすさレベル判定

まず、日本語を母語とする英語学習者の英語発話の書き起こしデータを英語母語話者が読み、各文の分かりやすさ（通じやすさ）を表1に示す3段階で判定した。発話内容は、自己紹介のほか、イラスト描写、ロールプレイ、コマ割マンガへのストーリー付けなどの task-based activity である。

表1 通じやすさのレベルと判定基準

レベル	判定基準
intelligible	文の意味を問題なく理解できる。誤りが含まれる場合もあるが、それも理解に支障をきたすことはない。
unclear	文の意味が全体的に、または部分的に不明確であったり、表現に不自然さが感じられる箇所があるが、少し推測すれば理解できる。
unintelligible	文の意味を全く理解できない。

その結果、3段階の通じやすさレベルの分布は表2のようになった。まず、調査対象である7,294文(47,486語)のうち、intelligibleと判定された文は5,774文(30,530語)で、全体の80%近くを占めた。また、全体の約17%である1,282文(15,058語)がunclearと判定された。unintelligibleと判定されたのは238文(2,198語)で、全体の約3%であった。また、一文あたりの平均語数(MLU:Mean Length of Utterance)は、unclearが11.74語と最も多く、中間がunintelligibleの9.23語で、最も少なかったのはintelligibleの5.28語であった。

表 2 通じやすさに関する調査対象データの内訳

レベル	文数(文)	語数(語)	MLU(語)
intelligible	5,774	30,530	5.28
unclear	1,282	15,058	11.74
unintelligible	238	2,198	9.23
合計	7,294	47,786	6.55

## 2.2. エラータグ付与～各誤りの特徴量抽出

次に、同じデータに対し、文中に誤りが含まれている場合は誤りを 47 種類に分類したエラータグ(和泉,内元,井佐原 2004)で標識付けを行った。最後に、3 段階の通じやすさレベルにおける各誤りの種類の特徴量を算出した。算出の基準は、「誤りのタイプ A の分かりやすさレベル I における特徴量 = 誤りのタイプ A が分かりやすさのレベル I のデータに出現する頻度 / 誤りのタイプ A がデータ全体に出現する頻度 (1,000 語で正規化)」とする。誤りのタイプ A がどの分かりやすさのレベルにおいて特徴的に出現するのを見ることにより、その誤りのタイプが文の通じやすさに及ぼす影響度を測るのが目的である。

表 3 誤りの各種類の通じやすさレベルにおける特徴量

誤りの種類		通じやすさのレベル		
		intelligible	unclear	unintelligible
形態素	(1)語形変化(名詞)	0.00	3.17	0.00
	(2)語形変化(動詞)	0.52	2.11	0.00
	(3)語形変化(形容詞)	0.78	1.58	0.00
	(4)名詞の可算性	0.60	1.95	0.00
文法	(5)名詞の数	0.71	1.61	0.75
	(6)主語・述語の人称・数的一致	0.68	1.69	0.64
	(7)動詞の時制	0.59	1.90	0.44
	(8)動詞の補部の形	0.41	2.23	0.62
	(9)副詞の位置	0.73	1.58	0.67
	(10)冠詞の選択	0.63	1.77	0.73
	(11)動詞の形	0.78	1.28	2.03
	(12)動詞の否定形	0.46	1.86	2.55
	(13)代名詞の照応先との数・人称の一致	0.32	0.93	1.82
語彙	(14)語彙選択(名詞)	0.51	1.85	1.93
	(15)語彙選択(動詞)	0.50	2.01	0.87
	(16)語彙選択(形容詞)	0.47	2.00	1.42
	(17)語彙選択(副詞)	0.52	1.75	2.40
	(18)語彙選択(非従属前置詞)	0.57	1.83	1.15
	(19)語彙選択(従属前置詞)	0.50	2.09	0.36
	(20)語彙選択(接続詞)	0.33	2.31	1.21
	(21)イディオム	0.50	1.97	1.15
	(22)和製英語	0.73	1.49	1.27
	その他	(23)語順	0.31	2.32
(24)大域的な誤り		0.33	2.09	2.78

注：セルの色が濃いほど特徴的に出現していることを表す。

その結果(表 3)のうち、本稿で対象とする単語レベルの語彙の誤りに関しては、各通じやすさレベルにも出現するものの、主に unclear な文に特徴的に出現することが分かった。

## 3. 語彙誤りと通じやすさの関係に関する調査

2.2 の調査結果を踏まえ、単語レベルの語彙誤りと通じやすさレベルの関係について、更に詳細な調査を行った。

### 3.1. 目的と手法

本調査では、単語レベルの語彙誤りのパターンが発話の通じやすさに及ぼす影響について大まかな傾向を把握することを目的とする。まず、先行研究で提案されている学習者言語における語彙の誤りパターンの細分類に基づき、2.の調査で用いられたデータ内の各語彙誤りがどのパターンに属するのか、人手で分類する。そして、各誤りパターンが 3 段階の通じやすさレベル(intelligible, unclear, unnatural)においてどのように分布するのかを調査する。

### 3.2. 調査のための準備:語彙誤りパターンの分類

学習者言語における語彙誤りパターンの細分類は、いくつかの先行研究によって試みられている(Engber 1995, Henning 1973, James 1998)。本調査では、James(1998)で提案されている分類基準のうち、書き言葉に特化した誤りを除いたものを更にアレンジして用いることにした。

表 4 に示されるのが James(1998)による細分類基準であるが、全体的に、「誤りの直接的な性質」を表す項目と「その性質を更に細分類した」項目と、「誤りが引き起こされる原因」を表す項目が混在しているようにとれる。例えば、語彙の誤りを直接的に(表層上の情報のみで)描写しようとする場合、最も公平な(分類者の主観が伴わない)方法は、誤用語と正用語が持つ概念にどのようなギャップがあるか、に基づいて細分類することである。つまり、産出された語が学習目標言語の語彙であり、一語である限り、すべての語彙誤りは、表 4 の(2),(7)-(11)のうちのいずれかに分類することができるはずである。また、例外的に母語をそのまま使用している場合は、(4)の「母語の使用」に分類される。一方、表 4 のその他の項目のうち、(12)-(14)は、その誤りがコロケーションに関わるものであった場合の細分類項目であり、(1)(3)(5)(6)は誤りの原因である。

表 4 語彙誤りの細分類 (James 1998)

形式上の誤り	(1)類似する形を持つ語間の誤用
	(2)同じ語の品詞違い
	(3>false friend
	(4)母語の使用
	(5)(母語に起因する)造語
	(6)母語からの直訳
意味上の誤り	(7)上位語の使用
	(8)下位語の使用
	(9)同位語間の誤用
	(10)同義語間の誤用
	(11)異なる語義を持つ語間の誤用
	(12)意味的制約が関わる語の組み合わせの誤用
	(13)慣用的な語の組み合わせの誤用
	(14)任意の語の組み合わせの誤用

そこで、表 4 を構成しなおし、かつ産出語が代名詞である場合、2 語以上である場合、造語である場合の誤りパターンを追加した上で、誤りの細分類の手順を整理すると図 1 のようになる。

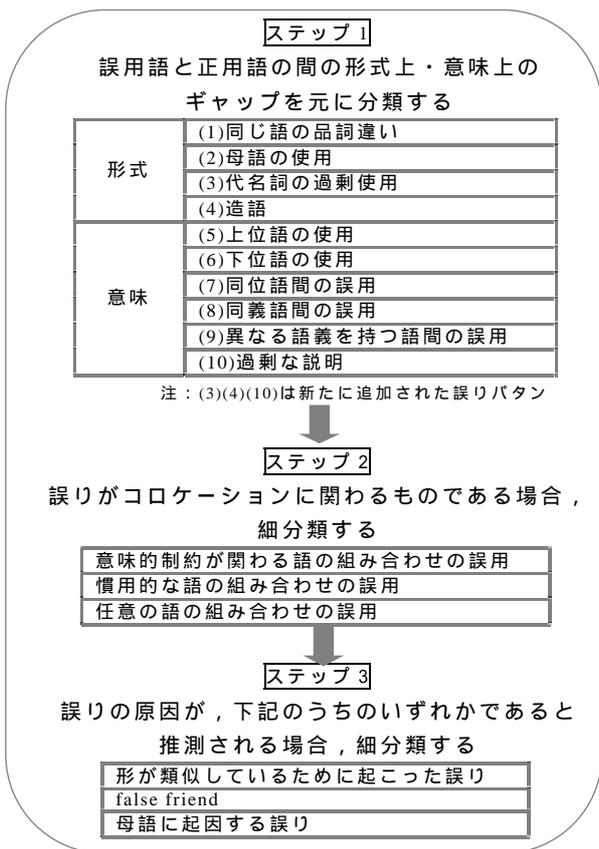


図 1 James(1998)の基準の再構成に基づく語彙誤りの細分類手順

まずステップ 1 として、誤用語と正用語の間の形式上または意味上のギャップを元に誤りを分類する。次に、ステップ 2 では、誤りがコロケーシ

ョンに関わるものである場合、更に細分類を行う。最後のステップ 3 で、誤りの原因が下記のうちのいずれかであると推測される場合、更に細分類する。

### 3.3. 調査対象

本調査では、名詞と動詞の単語レベルでの語彙誤りを対象に、ステップ 1 のみの分類を行った。それぞれの誤りが図 1 のステップ 1 で示されている(1)-(10)のいずれに属するかを手で判断した。(5)-(9)の判定の際は、WordNet3.0 で示されている概念階層を参照した。ステップ 1 のみを行う理由は、ステップ 2、ステップ 3 の細分類を行って包括的な結果を得られるだけのサンプル数がないことと、ステップ 3 の誤りの原因に関しては、図 1 に示されているものはごく一部にすぎないためである。

### 3.4. 調査結果と考察

調査結果は表 5 (名詞)と表 6 (動詞)のようになった。(1)-(10)の語彙誤りパターンが、intelligible, unclear, unintelligible と判定された文のグループにそれぞれ出現する頻度を示している。「%」の欄では、各通じやすさレベルでの語彙誤りの総数のうち、各パターンの誤りが占める割合を示している。全体のサンプル数が少なく、また、通じやすさレベル間にサンプル数の差があることから、この結果のみから一般化を行うことはできない。しかし、今後より定量的なデータで調査を行うにあたり、予備的な結果としての大まかな傾向を観察した。

表 5 名詞の語彙誤りパターンの分類の結果

	intelligible		unclear		unintelligible	
	頻度	%	頻度	%	頻度	%
(1)	10	19.6	8	8.1	0	0.0
(2)	5	9.8	3	3.0	1	7.1
(3)	0	0.0	4	4.0	2	14.3
(4)	4	7.8	5	5.1	1	7.1
(5)	4	7.8	11	11.1	1	7.1
(6)	4	7.8	16	16.2	2	14.3
(7)	8	15.7	6	6.1	2	14.3
(8)	11	21.6	8	8.1	4	28.6
(9)	4	7.8	26	26.3	1	7.1
(10)	1	2.0	12	12.1	0	0.0
計	51	100.0	99	100.0	14	100.0

表 6 動詞の語彙誤りパターンの分類の結果

	intelligible		unclear		unintelligible	
	頻度	%	頻度	%	頻度	%
(1)	5	8.6	12	9.7	2	25.0
(2)	0	0.0	0	0.0	0	0.0
(3)	-	-	-	-	-	-
(4)	0	0.0	1	0.8	0	0.0
(5)	6	10.3	4	3.2	2	25.0
(6)	1	1.7	1	0.8	0	0.0
(7)	6	10.3	6	4.8	1	12.5
(8)	25	43.1	37	29.8	0	0.0
(9)	13	22.4	62	50.0	3	37.5
(10)	2	3.4	1	0.8	0	0.0
計	58	100.0	124	100.0	8	100.0

たとえば、(5)-(9)はいずれも語彙の意味上の誤り (lexico-semantic errors)であるが、関連する概念を持つ語間の誤用である(5)-(8)と、異なる語義を持つ語間の誤用である(9)に関して、intelligible と unclear で比較すると、(5)-(8)は intelligible に出現する語彙誤りでの割合が高く、(9)は unclear に出現する語彙誤りでの割合が高い。また、同じ(9)に分類される誤りでも、intelligible に出現するものと unclear に出現するものとは性質に差が見られた。例えば、intelligible に出現する(9)の誤りの多くは、WordNet3.0 に基づく分類では、異なる語義を持つ語間での誤用として分類されてしまうが、全く無関連の語間での誤用ではなかった。例えば、“chicken” と言うべきところを、“cock-a-doodle” と言ったり、“food” と言うべき箇所を“eat” と言う例である。これらの語の組み合わせは、品詞も違うため、既存の概念階層において関連性を見出すことが難しいが、人間にとってはどちらかの語を聞いてもう一方の語を連想できる関係にあるといえる。Maera(1996)でも述べられているように、母語話者はある語とその語から連想される語の大きなネットワークを持っており、そのような知識は実際のコミュニケーションにおいて重要な役割を果たす。非母語話者にとっても、このような連想語ネットワークを広げ、適切な語が産出できない際の代用語として利用することは、非母語でのコミュニケーションにおける有効な communication strategy であると言える。これらの知見を踏まえて語彙教育の改善策について検討すると、新たな語彙を教示する際に、学習者がすでに知っている語彙との関係も同時に示すことにより複数の語彙が構造化されていであろうし、直接階層関係にない語同士でも、連想語としてネットワークが広がる。

unintelligible に関しては、今回対象としている単語レベルでの語彙誤りのサンプル数が極端に少ないため、統括的な知見は得られない。表3においても分かるとおり、文を unintelligible にする最も大きな要素は、単語レベルでの語彙誤りではなく単文では文法的にも語彙的にも誤りを含んでいないように見えるものの、文脈中で矛盾している、または論理的に誤っているものや、文構造を大きく変更しなければならない大域的な誤りが多く、局所的に誤りの検討を行うことがあまり意味を持たないことが分かった。

#### 4. まとめと今後の課題

本研究では、日本語を母語とする英語学習者の発話における語彙の誤りが文の通じやすさに及ぼす影響について調査を行った。単語レベルの語彙誤りは、通じやすさの中間レベルである unclear な文に最も特徴的に出現することが多く、

intelligible と unclear に出現する語彙の誤りには性質的な違いが見られた。intelligible に出現する語彙誤りは、語義の概念階層において上位、下位、または同位に属する語や同義語間での誤用が高い割合を占めた。一方、unclear では、異なる語義を持つ語間での誤用が最も高い割合を占めることがわかった。unintelligible における単語レベルの語彙誤りはサンプル数が極端に少なく、その性質を一般化することはできなかった。unintelligible に特徴的な誤りに関する詳細な分析は、今回のような局所的な誤りの特定ではなく、文全体の構造や複数の文の間の論理構造に関する大域的な調査が必要である。

本研究では、データに含まれる語彙誤りを個別に扱ったが、一文中に複数の誤りが含まれる例も多く、そのうちのどれが文の通じやすさにより大きな影響を与えているかはケースバイケースである。また、同じ種類に分類されている誤りでも、出現する文脈によって性質が異なることがある。今後の展望として、誤りを個別に扱うのではなく、異なる文脈での性質の変化や、複数の誤りの間の相互作用など、周辺情報と個々の誤りの関係も考慮に入れた分析を行いたい。

## 文 献

- Canale, M., and Swain, M. (1980) “Theoretical Bases of Communicative Approach to Second Language Teaching and Testing”, *Applied Linguistics*, 1, pp.1-47
- Engber, C. A. (1995) “The relationship of lexical proficiency to the quality of ESL composition”, *Journal of Second Language Writing*, vol.4, pp.139-155.
- Henning, G. H. (1973) “Remembering foreign language vocabulary: acoustic and semantic parameters”, *Language Learning*, vol.23, No.2, pp.185-197.
- James, C. (1998) *Errors in Language Learning and Use*, Longman, London, pp.142-153.
- Maera, P. (1996) “The dimensions of lexical competence”, In G. Brown, K. Malmkjar, and J. Williams (Eds.) (1996) *Performance & Competence in Second Language Acquisition*, Cambridge University Press, Cambridge, pp.35-53.
- 和泉絵美, 内元清貴, 井佐原均 (2004) 『日本人1200人の英語スピーキングコーパス』, アルク, 東京.